Junlin Ruan j2ruan@ucsd.edu PID: A17839687

BIMM-143: INTRODUCTION TO BIOINFORMATICS

The find-a-gene project assignment
http://thegrantlab.org/bimm143
Dr. Barry Grant

Version: 2025-04-07 (13:23:59 PDT on Mon, Apr 07)

Overview:

The find-a-gene project is a required assignment for BIMM-143. You should prepare a written report in **PDF** format that has responses to each question labeled **[Q1] - [Q10]** below. You may wish to consult the scoring rubric at the end of this document and the example report provided online (note that the example report is from a *previous quarter* and the questions may differ).

The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered in class.

Due Date:

Your responses to questions Q1-Q4 are due at 12pm on the **Monday of Week 5** (see the Assignments and Grading section of our website for details). Note that these first set of answers can be obtained very quickly (at best within 15 or 20 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due at 12pm on the **Monday of Week 10**.

Submission instructions:

Your report formatted as a **PDF document** should be uploaded to ***GradeScope***. Please make sure to include your UCSD email and PID number on the first page.

Be sure to include your UCSD email and PID number on the first page of your report.

Submit your preliminary report with answers to Q1-Q4 as soon as you can so we can determine if you have found a novel gene. Submit this preliminary report as one document with screen shots of the results inserted appropriately.

See the demonstration report linked to on the course website for an example of format. Note again that example questions may differ. I will indicate on GradeScope my decision (1pt indicating all is good, 0pts revisions required). You should proceed with subsequent questions only after we are sure you have found a novel gene (and thus be successful in the later stages of the project).

For the final report add your results for Q5-Q10 to the preliminary report and submit the final

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

document containing your results for all questions.
Please do not send only Q5-Q10 answers as the final report.

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

Questions:

[Q1] Tell me the name of a protein you are interested in. Include the species, accession number and known function. This can be a human protein or a protein from any other species as long as it's function is known.

If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

RET4_HUMAN: the gene name is RBP4, in Homo sapiens; the primary accession number is P02753. It is a retinol-binding protein that mediates retinol transport in blood plasma, from the liver stores to the peripheral tissues.

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called `Screen Shot [].png` in your Desktop directory). It is **not** necessary to print out all of the blast results if there are many pages.

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

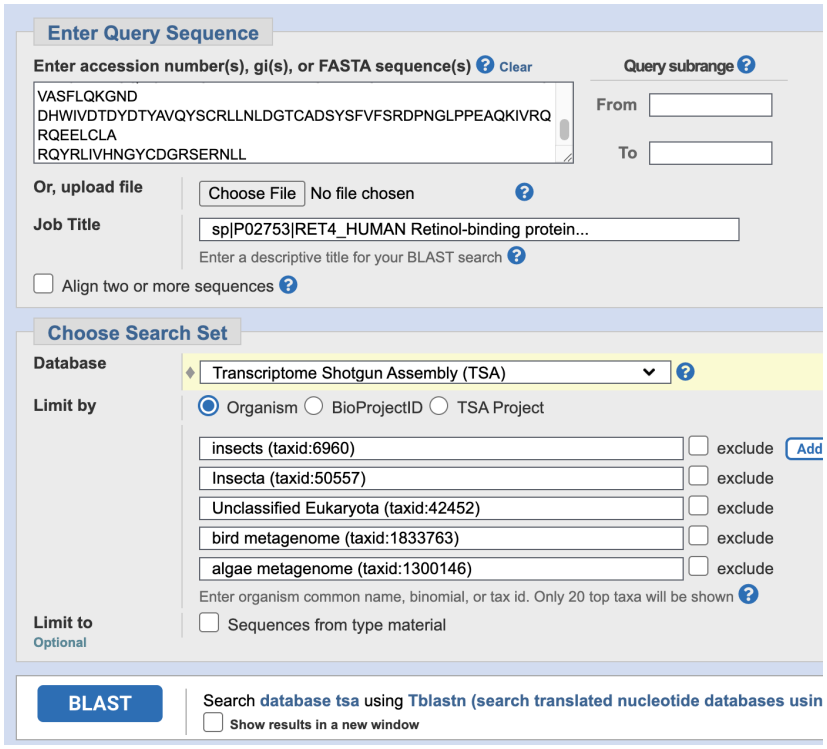
In general, [Q2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of [Q4]), and a non-homologous result.

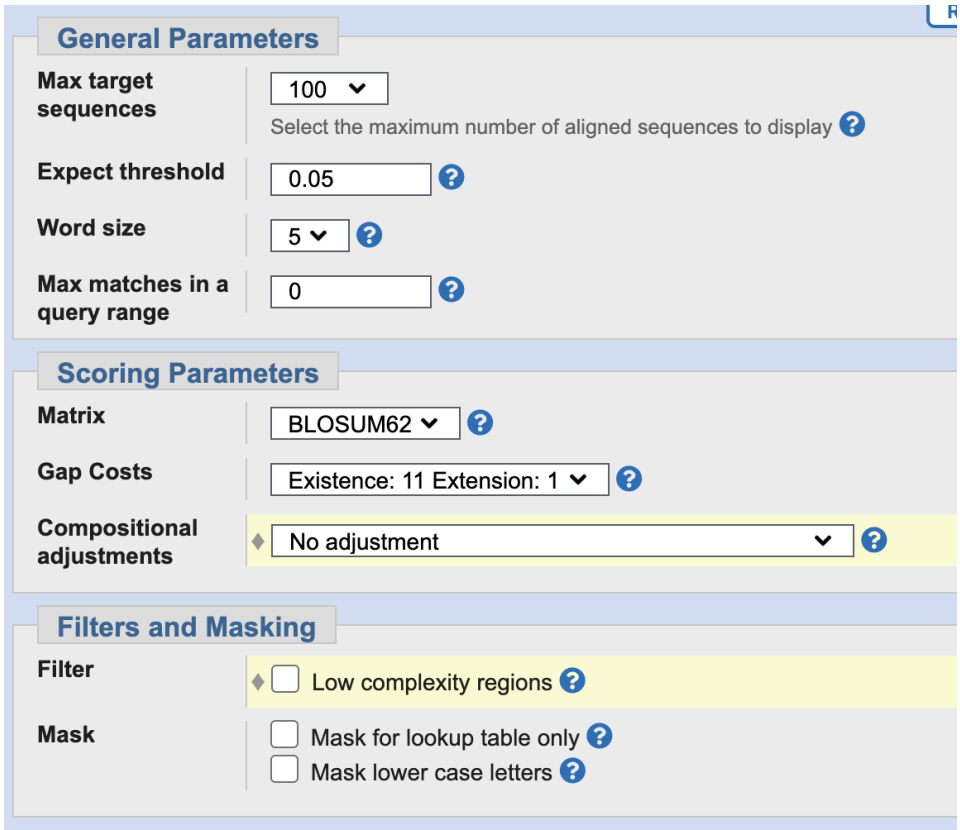
If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

A tblastn search was performed using the RET4_HUMAN protein as the query. The database selected was “Transcriptome Shotgun Assembly (TSA)” to look for genomic DNA sequences. The search was limited to the genomes of insects, unclassified eukaryotes, and the metagenomes of birds and algae to identify homologous but novel genomic sequences. Unchecked the filter and compositional adjustments to increase

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

the chance of novelty.





Junlin Ruan j2ruan@ucsd.edu PID: A17839687

Below is the blast result. A match that is homologous to the query protein that is likely novel is Select seq gb|GJZQ01401365.1| Bactericera cockerelli TRINITY_DN256384_c0_g1_i1 373 373 100% 5e-129 86.07% 813 GJZQ01401365.1

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Zorotypus shannoni isolate wild CL526.Contig1_ZOR_Zsp_MAN	395	395	100%	1e-137	92.04%	803	GJWI01029312.1
Zorotypus shannoni isolate wild CL526.Contig2_ZOR_Zsp_MAN	395	790	100%	2e-136	92.04%	1484	GJWI01036097.1
Bactericera cockerelli TRINITY_DN256384_c0_g1_i1	373	373	100%	5e-129	86.07%	813	GJZQ01401365.1
TSA: Serangium japonicum TRINITY_DN41571_c0_g1 transcribed RNA sequence	273	273	95%	3e-89	66.15%	990	GGMU01010997.1
Trichoplusia ni TRINITY_DN216245_c0_g1_i1	248	248	96%	1e-78	56.99%	1261	GHOK01008465.1
Trichoplusia ni TRINITY_DN18776_c0_g1_i1	211	211	93%	3e-64	50.00%	1088	GHOK01128344.1
Trichoplusia ni TRINITY_DN46851_c0_g1_i1	203	203	85%	8e-62	52.02%	829	GHOK01040714.1
Nezara viridula TRINITY_DN116076_c0_g1_i1	192	192	44%	6e-58	98.88%	268	GIBW01247166.1
TSA: Oropsylla silantiewi comp652932_c0_seq1 transcribed RNA sequence	184	184	61%	9e-55	62.60%	402	GAWY01035733.1
TSA: Loxostege sticticalis c81754.graph_c0 transcribed RNA sequence	134	134	40%	5e-35	69.14%	245	GFCJ01067156.1
TSA: Callosobruchus maculatus TR42114-c0_g1_i1 transcribed RNA sequence	108	108	38%	4e-25	55.84%	232	GEUD01191601.1
TSA: Mastotermes darwiniensis C430437_a_3_0_I_410 transcribed RNA sequence	97.1	97.1	25%	2e-20	86.27%	410	GAZE02106792.1
TSA: Oropsylla silantiewi comp540088_c0_seq1 transcribed RNA sequence	71.6	71.6	25%	1e-10	62.00%	217	GAWY01032403.1
Trichoplusia ni TRINITY_DN31835_c0_g1_i1	68.6	68.6	29%	2e-09	57.63%	251	GHOK01064199.1
TSA: Carposina sasakii c86671.graph_c0 transcribed RNA sequence	62.4	62.4	24%	5e-07	51.02%	317	GGMY01084648.1
TSA: putative bacterial lipocalin cell envelope bioproteinis lipocalin, complete cds	62.8	62.8	74%	7e-07	29.68%	603	GGFM01005965.1
TSA: Extatosoma tiaratum comp45074_c0_seq1 transcribed RNA sequence	62.4	62.4	67%	1e-06	31.47%	694	GAWG01015524.1
TSA: putative bacterial lipocalin cell envelope bioproteinis, complete cds	61.2	61.2	89%	2e-06	28.57%	600	GGFK01011687.1
Extatosoma tiaratum breed wildtype s2345_L_5061_0_a_46_8_I_1436	63.5	63.5	67%	3e-06	32.17%	1436	GDZM01038056.1
TSA: Anopheles aquasalis megaclu_asbSigP-16891 mRNA sequence	60.8	60.8	85%	4e-06	28.16%	630	GAMD01000603.1

TSA: Bactericera cockerelli TRINITY_DN256384_c0_g1_i1, transcribed RNA sequence

Sequence ID: GJZQ01401365.1 **Length:** 813 **Number of Matches:** 1

Range 1: 31 to 633 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
373 bits(958)	5e-129	173/201(86%)	190/201(94%)	0/201(0%)	+1

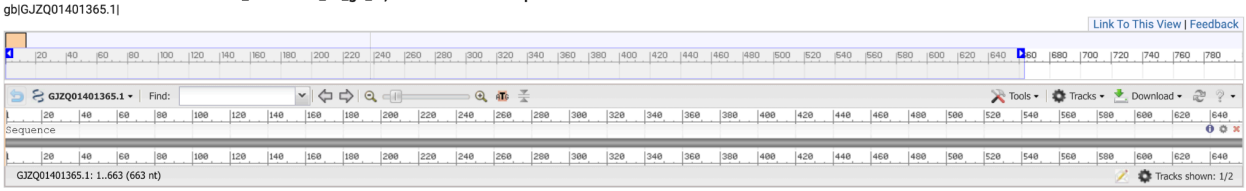
```
Query  1    MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  60
            M+WVWAL+LLAALGSG AERDCRVS+FRVKENFDKARFSGTWYA+AKKDPEGLFLQDNI+
Sbjct  31   MEWVWALVLLAALGSGWAERDCRVSNFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII  210

Query  61   AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  120
            AEFSVDE GQMSATAKGRVRLL+NW+VCADMVGTFTDTEDPAKFKMKYWGVASFLQ+GND
Sbjct  211  AEFSVDENGQMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND  390

Query  121  DHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA  180
            DHWI+DTDYDT+A+QYSCRLLNLDGTCADSYSFVFSR P+GL PEA+++VRQRQEELCL
Sbjct  391  DHWIIDTDYDTFALQYSCRLLNLDGTCADSYSFVFSRHPSGLTPEARRLVRQRQEELCLD  570

Query  181  RQYRLIVHNGYCDGRSERNLL  201
            RQYR I HNGYC  +  RN+L
Sbjct  571  RQYRWIEHNGYCQSKLSRNIL  633
```

TSA: Bactericera cockerelli TRINITY_DN256384_c0_g1_i1, transcribed RNA sequence
gb|GJZQ01401365.1|



Junlin Ruan j2ruan@ucsd.edu PID: A17839687

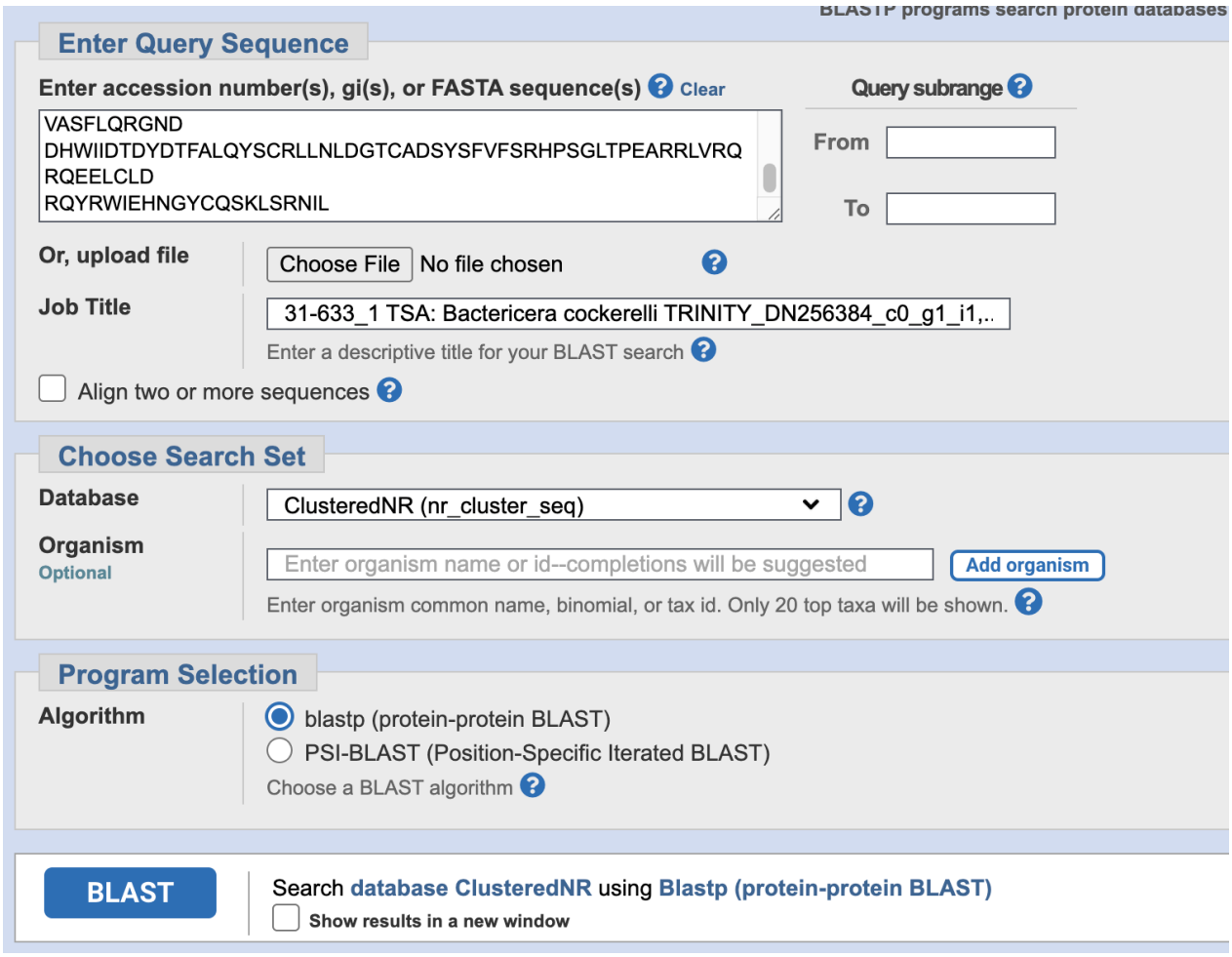
[Q3] Gather information about this "novel" **protein**. At a minimum, show me the protein sequence of the "novel" protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don't forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don't have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format. Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

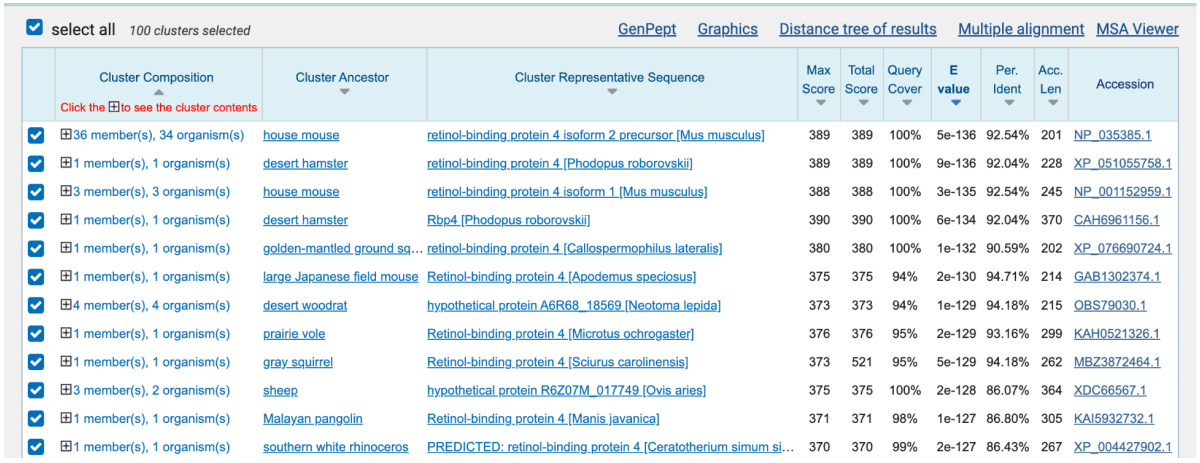
```
>31-633_1 TSA: Bactericera cockerelli TRINITY_DN256384_c0_g1_i1,
transcribed RNA sequence
MEWVWALVLLAALGSGWAERDCRVSNFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII
AEFSVDENGQMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND
DHWIIDTDYDTFALQYSCRLLNLDGTCADSYSFVFSRHPSGLTPEARRLVRQRQEELCLD
RQYRWIEHNGYCQSKLSRNIL
```

[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as "unknown"). Someone has already found and annotated this sequence, and assigned it an accession number.
- If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.
- If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene.
- If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.

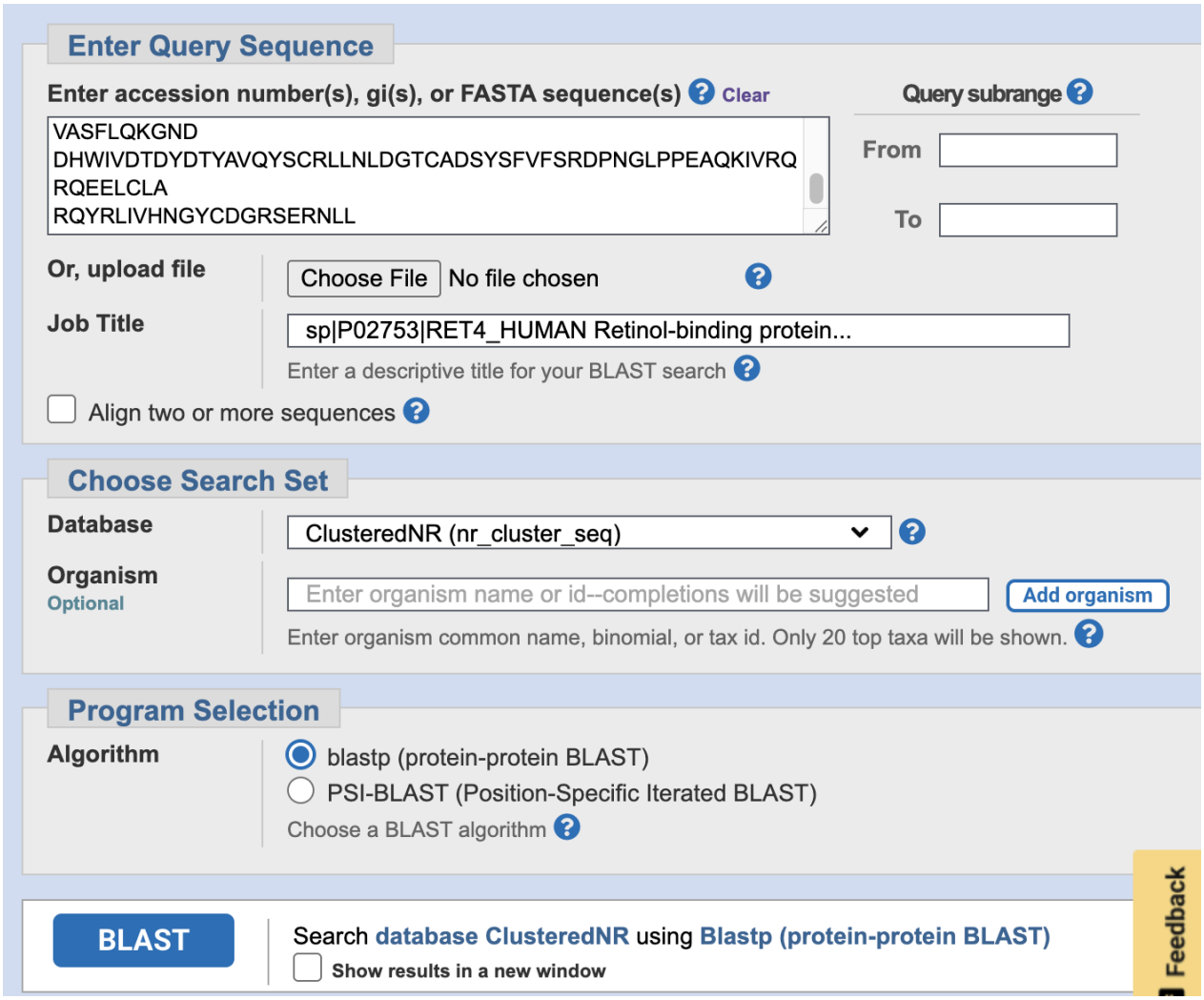
Junlin Ruan j2ruan@ucsd.edu PID: A17839687





Cluster Composition (Click the ⊞ to see the cluster contents)	Cluster Ancestor	Cluster Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
⊞36 member(s), 34 organism(s)	house mouse	retinol-binding protein 4 isoform 2 precursor [Mus musculus]	389	389	100%	5e-136	92.54%	201	NP_035385.1
⊞1 member(s), 1 organism(s)	desert hamster	retinol-binding protein 4 [Phodopus roborovskii]	389	389	100%	9e-136	92.04%	228	XP_051055758.1
⊞3 member(s), 3 organism(s)	house mouse	retinol-binding protein 4 isoform 1 [Mus musculus]	388	388	100%	3e-135	92.54%	245	NP_001152959.1
⊞1 member(s), 1 organism(s)	desert hamster	Rbp4 [Phodopus roborovskii]	390	390	100%	6e-134	92.04%	370	CAH6961156.1
⊞1 member(s), 1 organism(s)	golden-mantled ground sq...	retinol-binding protein 4 [Callospermophilus lateralis]	380	380	100%	1e-132	90.59%	202	XP_076690724.1
⊞1 member(s), 1 organism(s)	large Japanese field mouse	Retinol-binding protein 4 [Apodemus speciosus]	375	375	94%	2e-130	94.71%	214	GAB1302374.1
⊞4 member(s), 4 organism(s)	desert woodrat	hypothetical protein A6R68_18569 [Neotoma lepida]	373	373	94%	1e-129	94.18%	215	OBS79030.1
⊞1 member(s), 1 organism(s)	prairie vole	Retinol-binding protein 4 [Microtus ochrogaster]	376	376	95%	2e-129	93.16%	299	KAH0521326.1
⊞1 member(s), 1 organism(s)	gray squirrel	Retinol-binding protein 4 [Sciurus carolinensis]	373	521	95%	5e-129	94.18%	262	MBZ3872464.1
⊞3 member(s), 2 organism(s)	sheep	hypothetical protein R6Z07M_017749 [Ovis aries]	375	375	100%	2e-128	86.07%	364	XDC66567.1
⊞1 member(s), 1 organism(s)	Malayan pangolin	Retinol-binding protein 4 [Manis javanica]	371	371	98%	1e-127	86.80%	305	KAI5932732.1
⊞1 member(s), 1 organism(s)	southern white rhinoceros	PREDICTED: retinol-binding protein 4 [Ceratotherium simum si...	370	370	99%	2e-127	86.43%	267	XP_004427902.1

Junlin Ruan j2ruan@ucsd.edu PID: A17839687



1 member(s), 1 organism(s)	Iberian mole	retinol-binding protein 4 [Talpa occidentalis]	386	386	100%	5e-135	88.56%	201	XP_037352753.1
1 member(s), 1 organism(s)	NA	Retinol-binding protein 4 [Galemys pyrenaicus]	390	390	99%	8e-135	90.91%	308	KAG8523527.1
2 member(s), 2 organism(s)	southern two-toed sloth	retinol-binding protein 4 [Choloepus didactylus]	384	384	100%	4e-134	87.56%	201	XP_037660868.1
1 member(s), 1 organism(s)	Cape elephant shrew	PREDICTED: retinol-binding protein 4 [Elephantulus edwa...	382	382	100%	3e-133	88.06%	201	XP_006880416.1
1 member(s), 1 organism(s)	NA	hypothetical protein HPG69_019583 [Diceros bicornis min...	384	384	99%	1e-132	86.60%	302	KAF5911215.1
1 member(s), 1 organism(s)	nine-banded armadillo	retinol-binding protein 4 isoform X1 [Dasypus novemcinctus]	382	382	100%	2e-132	89.05%	283	XP_012376724.2
1 member(s), 1 organism(s)	muntjak	hypothetical protein FD754_011836 [Muntiacus muntjak]	377	377	100%	7e-131	88.56%	233	KAB0346979.1
12 member(s), 11 organism(s)	Southern elephant seal	hypothetical protein GH733_001004 [Mirounga leonina]	375	375	95%	9e-131	92.11%	198	KAF3827769.1
1 member(s), 1 organism(s)	sheep	retinol-binding protein 4 isoform X1 [Ovis aries]	375	375	95%	7e-130	92.63%	257	XP_060260599.1
1 member(s), 1 organism(s)	Yarkand deer	hypothetical protein G4228_010897 [Cervus hanglu yarka...	374	374	95%	9e-130	92.11%	229	KAF4019100.1
1 member(s), 1 organism(s)	NA	retinol binding protein 4 [Myotis myotis]	373	373	98%	2e-129	88.83%	232	KAF6304270.1
1 member(s), 1 organism(s)	goat	PREDICTED: retinol-binding protein 4 isoform X1 [Capra ...	372	372	94%	3e-129	92.55%	225	XP_017896583.1
1 member(s), 1 organism(s)	Sunda flying lemur	PREDICTED: retinol-binding protein 4 [Galeopterus varieg...	374	374	96%	1e-128	91.19%	304	XP_008587057.1
36 member(s), 34 organism(s)	house mouse	retinol-binding protein 4 isoform 2 precursor [Mus musculus]	369	369	100%	3e-128	85.57%	201	NP_035385.1
3 member(s), 3 organism(s)	house mouse	retinol-binding protein 4 isoform 1 [Mus musculus]	369	369	100%	1e-127	85.57%	245	NP_001152959.1

It is a novel but still homologous gene.

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

[Q5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.

Re-labeled sequences for alignment:

```
>Human_RBP4 Retinol-binding protein 4 [Homo sapiens]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND
DHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA
RQYRLIVHNGYCDGRSERNLL

>Bcockerelli_RBP_like TRINITY_DN256384_c0_g1_i1 Retinol-binding
protein-like [Bactericera cockerelli]
MEWVWALVLLAALGSGWAERDCRVSNFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII
AEFSVDENGQMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND
DHWIIDTDYDTFALQYSCRLLNLDGTCADSYSFVFSRHPSGLTPEARRLVRQRQEELCLD
RQYRWIEHNGYCQSKLSRNIL

>Mouse_RBP4 Retinol-binding protein 4 isoform 2 precursor [Mus musculus]
MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDPEGLFLQDNIIAEFSVDEKGHMS
ATAKGRVR
LLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRLQNLDGTCADSYS
FVFSRDPN
GLSPETRRLVRQRQEELCLERQYRWIEHNGYCQSRPSRNSL

>Phodopus_RBP4 Retinol-binding protein 4 [Phodopus roborovskii]
MELRAAQSSCPAVTPRLSRRADSCACEMEWVWALVLLAVLGGGSAERDCRVSSFRVKENFDKARFSGIWYAI
```

AKKDPEGL

FLQDNIIAEFSVDEKGHMSATAKGRVRILSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWII
DTDYDTFA

LQYSCRLQNLDGTCADSYSFVFSRDPNGLTPETRRLVRQRQEELCLERQYRWIEHNGYCQSRPSRNSL

>Callospermophilus_RBP4 Retinol-binding protein 4 [Callospermophilus
lateralis]

MEWVWALVVLLAALGSGRAERDCRVSSFRVKENFDKTLFSGTWYAIAKKDPEGLFLQDNIVAEFSVDENGHM
SATAKGRV

RLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGIDDHWIIDTDYHTFALQYSCRLLNFDGTCADSY
SFVFARDP

NGLTPEIRKLVRQRQEELCLDRQYRWIEHNGYCQSKTGENLL

>Apodemus_RBP4 Retinol-binding protein 4 [Apodemus speciosus]

MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDPEGLFLQDNIIAEFSVDEKGQMS
ATAKGRVR

LLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTYALQYSCRLQNLDGTCADSYS
FVFSRDPS

GLTPETRRLVRQRQEELCLERQYRWIEHNVKVTVKAYPRETVCSDVKDVKFENF

>Microtus_RBP4 Retinol-binding protein 4 [Microtus ochrogaster]

MGRPTARSSCLAVTRALVPQGGLLSEMEWVWALVLLAALGGSSAERDCRVSSFRVKENFDNARFSGLWYAIA
KKDPEGLF

LQDNIIAEFSVDEKGHMSATAKGRVRILSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIID
TDYDTFAL

QYSCRLLNLDGTCADSYSFVFSRDPNGLTPETRRLVRQRQEELCLERQYRWIEHNGPIDFSSFKQLRTLSPF
LVVFLRIC

ASPRSQSQSAHGSGLLLQTSGAAHRTLHELLMLACQALPDLALESHWHPSMGTRAAPLS

>Manis_RBP4 Retinol-binding protein 4 [Manis javanica]

MFPNPLHNRASEGGGAAGGIADALSPWEATAVSQADGSTEPERAGPEGPPRPQLAAAGTSAGEPGQGLFPRI
EATPHGSP

TEPRAALEVRLSPKRTLPAPRRRADSGEMEWVWALVLLAALGSARAERDCRVSSFRVKENFDKTRFSGTWYA
MAKKDPEG

LFLQDNIIAEFSVDESGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWI
IDTDYDTY

AVQYSCRLLNLDGTCADSYSFVFARNPNGLPPEVQKIVRQRQEELCLARQYRLIMHNGYCDGRLA

>Choloepus_RBP4 Retinol-binding protein 4 [Choloepus didactylus]

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

MEWVWALVLLAALGSSRAERDCRVSTFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIIAEFSVDENGQMT
ATAKGRVR

LFNNWDVCADMVGTFTDTEDPAKFKMKYWGAASFLQKGYDDHWIIDTDYDTYAVQYACRLQNLDGTCADSYS
FVFARDPQ

GLPPEVQRVVRQRQEELCLGRQYRLIVHDGYCNSKSEKNVL

Alignment:

Obtained using MUSCLE (version 3.8) at EBI:

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)



Callospermophilus_RBP4
------------------------------------------------------------

Microtus_RBP4
------------------------------------------------------------

Apodemus_RBP4
------------------------------------------------------------

Bcockerelli_RBP_like
------------------------------------------------------------

Mouse_RBP4
------------------------------------------------------------

Phodopus_RBP4
------------------------------------------------------------

Choloepus_RBP4
------------------------------------------------------------

Human_RBP4
------------------------------------------------------------

Manis_RBP4
MFPNPLHNRASEGGGAAGGIADALSPWEATAVSQADGSTEPERAGPEGPPRPQLAAAGTS



Callospermophilus_RBP4
--------------------------------------------------MEWVWALVVLLA

Microtus_RBP4
---------------------MGRPTARSSCLAVTRALVPQGGLLSE-MEWVWAL-VLLA

Apodemus_RBP4
--------------------------------------------------MEWVWAL-VLLA

Bcockerelli_RBP_like
```

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

```
--------------------------------------------------MEWVWAL-VLLA

Mouse_RBP4
--------------------------------------------------MEWVWAL-VLLA

Phodopus_RBP4
---------------------MELRAAQSSCPAVTPRLSRRADSCACEMEWVWAL-VLLA

Choloepus_RBP4
--------------------------------------------------MEWVWAL-VLLA

Human_RBP4
--------------------------------------------------MKWVWAL-LLLA

Manis_RBP4
AGEPGQGLFPRIEATPHGSPTEPRAALEVRLSPKRTLPAPRRRADSGEMEWVWAL-VLLA

*:***** :***


Callospermophilus_RBP4
ALGSGRAERDCRVSSFRVKENFDKTLFSGTWYAIAKKDPEGLFLQDNIVAEFSVDENGHM

Microtus_RBP4
ALGGSSAERDCRVSSFRVKENFDNARFSGLWYAIAKKDPEGLFLQDNIIAEFSVDEKGHM

Apodemus_RBP4
ALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDPEGLFLQDNIIAEFSVDEKGQM

Bcockerelli_RBP_like
ALGSGWAERDCRVSNFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNIIAEFSVDENGQM

Mouse_RBP4
ALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDPEGLFLQDNIIAEFSVDEKGHM

Phodopus_RBP4
VLGGGSAERDCRVSSFRVKENFDKARFSGIWYAIAKKDPEGLFLQDNIIAEFSVDEKGHM

Choloepus_RBP4
ALGSSRAERDCRVSTFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIIAEFSVDENGQM

Human_RBP4
ALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQM

Manis_RBP4
ALGSARAERDCRVSSFRVKENFDKTRFSGTWYAMAKKDPEGLFLQDNIIAEFSVDESGQM

                             .**.. ********.********:: ***
***:***************:*******.*:*


Callospermophilus_RBP4
SATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGIDDHWIIDTDYHT

Microtus_RBP4
SATAKGRVRILSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDT

Apodemus_RBP4
SATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDT
```

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

```
Bcockerelli_RBP_like
SATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDT

Mouse_RBP4
SATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDT

Phodopus_RBP4
SATAKGRVRILSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDT

Choloepus_RBP4
TATAKGRVRLFNNWDVCADMVGTFTDTEDPAKFKMKYWGAASFLQKGYDDHWIIDTDYDT

Human_RBP4
SATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDT

Manis_RBP4
SATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDT

:********::.**:************************.*****.* *****:**** *


Callospermophilus_RBP4
FALQYSCRLLNFDGTCADSYSFVFARDPNGLTPEIRKLVRQRQEELCLDRQYRWIEHN--

Microtus_RBP4
FALQYSCRLLNLDGTCADSYSFVFSRDPNGLTPETRRLVRQRQEELCLERQYRWIEHNGP

Apodemus_RBP4
YALQYSCRLQNLDGTCADSYSFVFSRDPSGLTPETRRLVRQRQEELCLERQYRWIEHNVK

Bcockerelli_RBP_like
FALQYSCRLLNLDGTCADSYSFVFSRHPSGLTPEARRLVRQRQEELCLDRQYRWIEHN--

Mouse_RBP4
FALQYSCRLQNLDGTCADSYSFVFSRDPNGLSPETRRLVRQRQEELCLERQYRWIEHN--

Phodopus_RBP4
FALQYSCRLQNLDGTCADSYSFVFSRDPNGLTPETRRLVRQRQEELCLERQYRWIEHN--

Choloepus_RBP4
YAVQYACRLQNLDGTCADSYSFVFARDPQGLPPEVQRVVRQRQEELCLGRQYRLIVHD--

Human_RBP4
YAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHN--

Manis_RBP4
YAVQYSCRLLNLDGTCADSYSFVFARNPNGLPPEVQKIVRQRQEELCLARQYRLIMHN--

                              :*:**:*** *:************:* *.**.**
..:********** **** * *:


Callospermophilus_RBP4
----GYCQSKTGENLL--------------------------------------------

Microtus_RBP4
IDFSSFKQLRTLSPFLVVFLRICASPRSQSQSAHGSGLLLQTSGAAHRTLHELLMLACQA

Apodemus_RBP4
```

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

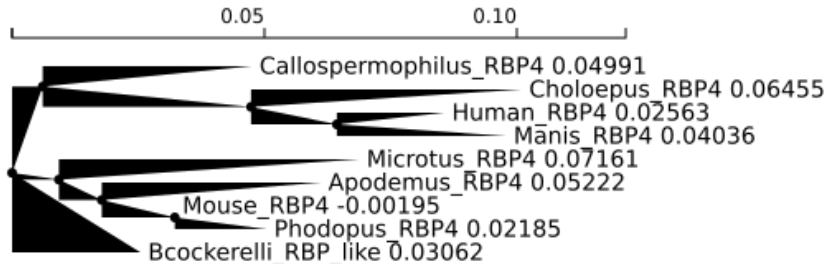
```
VTVKAYPRETVCSDVKDVKFENF------------------------------------
Bcockerelli_RBP_like
----GYCQSKLSRNIL--------------------------------------------
Mouse_RBP4
----GYCQSRPSRNSL--------------------------------------------
Phodopus_RBP4
----GYCQSRPSRNSL--------------------------------------------
Choloepus_RBP4
----GYCNSKSEKNVL--------------------------------------------
Human_RBP4
----GYCDGRSERNLL--------------------------------------------
Manis_RBP4
----GYCDGRLA------------------------------------------------
                                  .:

Callospermophilus_RBP4      ----------------------
Microtus_RBP4               LPDLALESHWHPSMGTRAAPLS
Apodemus_RBP4               ----------------------
Bcockerelli_RBP_like        ----------------------
Mouse_RBP4                  ----------------------
Phodopus_RBP4               ----------------------
Choloepus_RBP4              ----------------------
Human_RBP4                  ----------------------
Manis_RBP4                  ----------------------
```

[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

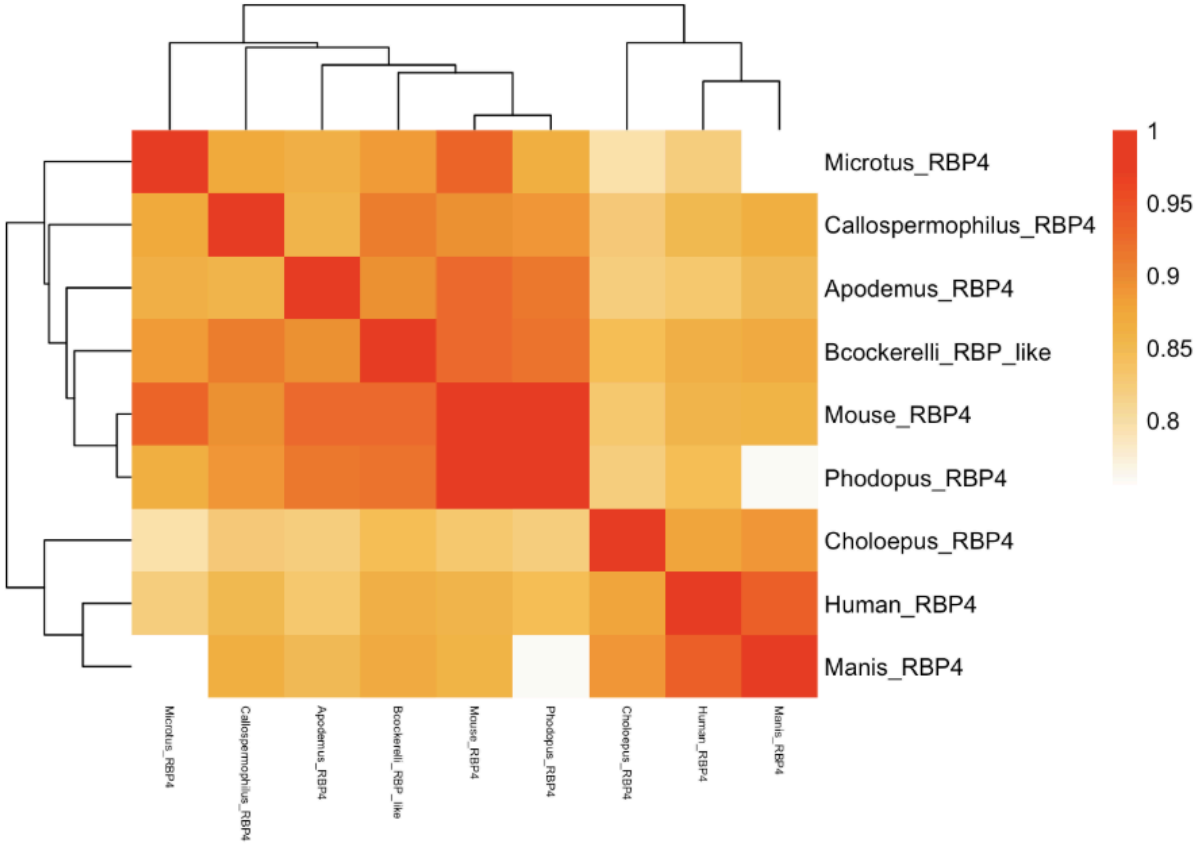
Imported the sequence with Phylip, align with MUSCLE, and create a neighbor-joining tree:

Junlin Ruan j2ruan@ucsd.edu PID: A17839687



[Q7] Generate a sequence identity based **heatmap** of your aligned sequences using R.

If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and "Save as" FASTA format for example). Read this FASTA format alignment into R with the help of functions in the **Bio3D package**. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.



Junlin Ruan j2ruan@ucsd.edu PID: A17839687

[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.

List the top 3 *unique* hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above.

Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

ID	Technique	Resolution	Source	E-value	%Identity
1HBQ	X-RAY DIFFRACTION	1.70 Å	Bos taurus	1e-122	86.89%
4O9S	X-RAY DIFFRACTION	2.30 Å	Homo sapiens	2e-122	85.48%
3FMZ	X-RAY DIFFRACTION	2.90 Å	Homo sapiens	3e-122	85.48%

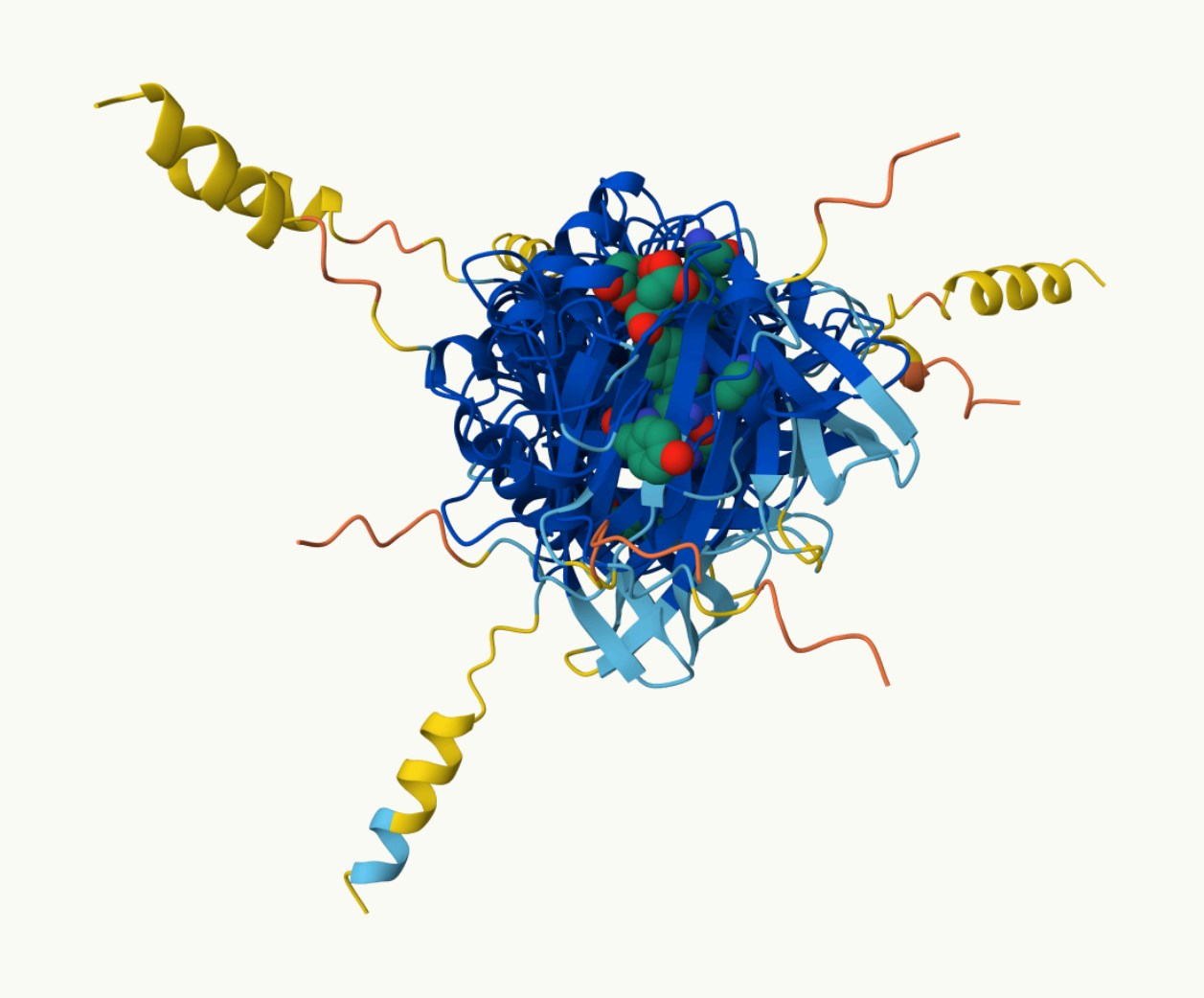
[Q9] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.

Note that this can take some time depending upon your sequence length. If your

model is taking many hours to generate or your input sequence yields a “too many amino acids” (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches.

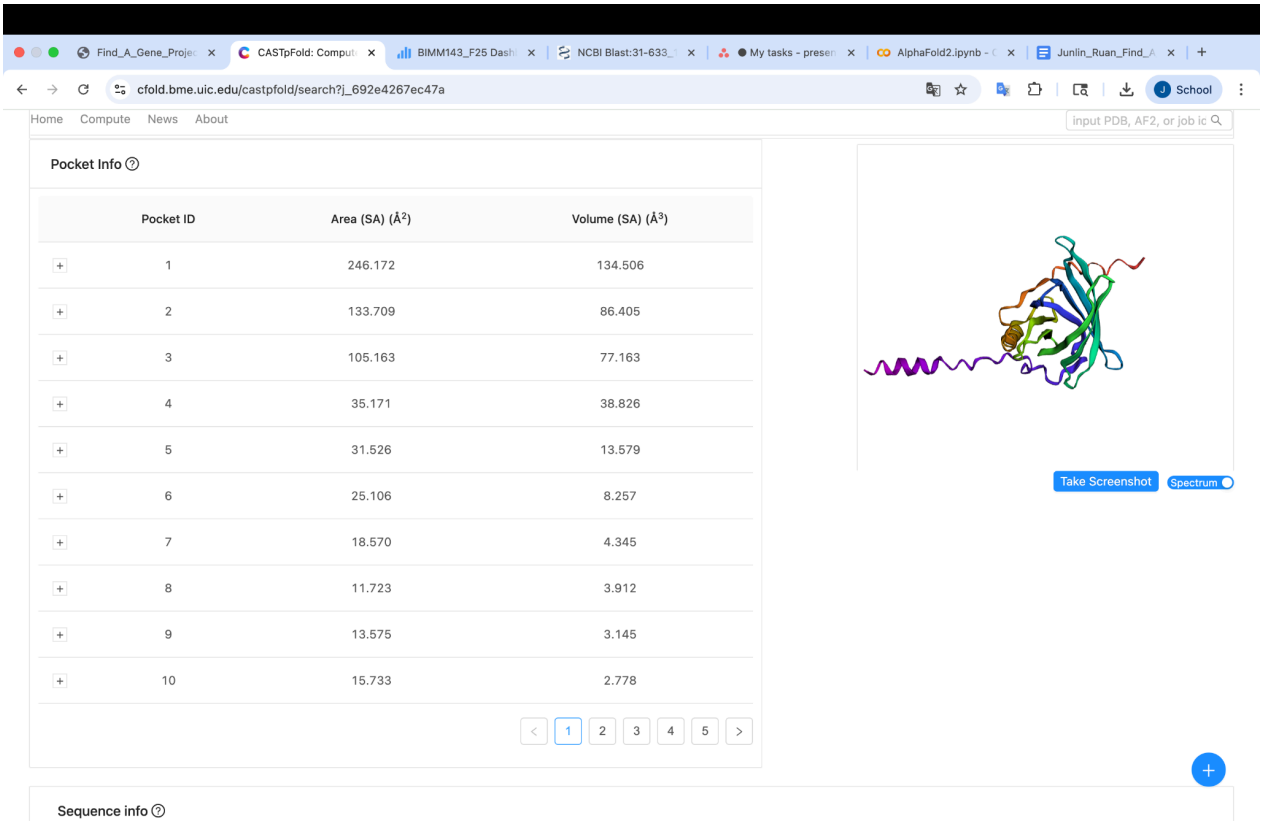
Once complete save the resulting PDB format file for your records. Finally, generate a molecular figure of your generated PDB structure using the **Mol* viewer** online (or VMD/PyMol/Chimera if you prefer). To complete your analysis you should highlight *conserved residues* that are likely to be functional as **spacefill** and the protein as **cartoon** colored by local alpha fold *pLDDT quality score*. You can determine conserved residues from the alignment generated by the AlphaFold server and use a conservation cutoff appropriate for the diversity of your protein alignment (e.g. between 60% and 99% conserved). Note that *pLDDT* score is contained in the B-factor column of your PDB downloaded file. Please use a white or transparent background for your figure (i.e. not the default black in PyMol/VMD/Chimera etc.).

Junlin Ruan j2ruan@ucsd.edu PID: A17839687



[Q10] (i) Using your computed structure model (or your closest homologue of known structure from the PDB) predict and locate potential small molecule binding sites using the CASTpFold server (https://cfold.bme.uic.edu/castpfold/). Provide an image or screen-shot of your largest predicted pockets “negative volume” and provide it’s **area** and **volume**.

Junlin Ruan j2ruan@ucsd.edu PID: A17839687



The largest pocket identified was Pocket 1, with an area of 246.172 Å² and a volume of 134.506 Å³.

(ii) Perform a “Target” search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any **Target Associated Assays** and **ligand efficiency data** reported that may be useful starting points for exploring potential inhibition of your novel protein? If there are no assays listed here simply list “non available as of [date]”.

Non available as of Dec. 1st, 2025.
https://www.ebi.ac.uk/chembl/search_results/GJZQ01401365.1

Junlin Ruan j2ruan@ucsd.edu PID: A17839687

(iii) Briefly discuss (100 words max) the **druggability** of your novel protein based on:

- Presence of well-defined pockets (output of tools like CASTpFold), - Existence of known inhibitors for related proteins (your search of ChEMBEL), - Conservation of binding sites across homologs (your conservation analysis in Q10),
- Potential therapeutic applications if this protein were targeted (you can use ChatGPT, Claude etc. backed up by your reading of the literature here).

The novel protein contains a well-defined, largest pocket (134.51 Å^3), which is compatible with small-molecule binding. However, the ChEMBL search returned no target-associated binding data or known inhibitors for this sequence (as of December 1st, 2025). Only eight residues show conservation ≥45%, with the highest at 54.69%, suggesting limited evolutionary constraint and potential selectivity challenges. Despite this, the lack of existing chemical matter indicates that, if future research links this protein to a disease mechanism, inhibition could establish a new therapeutic strategy. In such a case, any successful ligand would represent a first-in-class drug candidate.